PATENT ABSTRACTS OF JAPAN

(11)Publication number : 10-097286
(43)Date of publication of application : 14.04.1998

(51)Int.CI. G10L 3/00
G10L 3/00
G06F 17/28

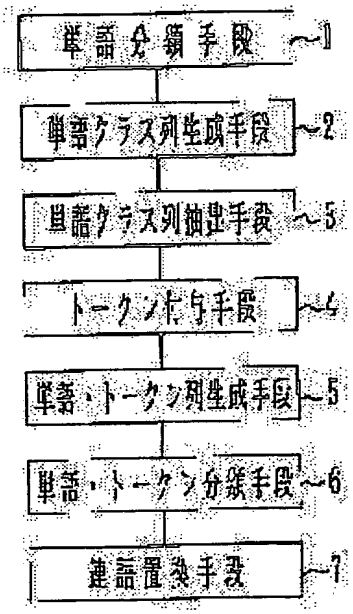
(21)Application number : 09-167243
(22)Date of filing : 24.06.1997

(71)Applicant : FUJITSU LTD
(72)Inventor : SHIODA AKIRA

(30)Priority
Priority number : 08204986 Priority date : 02.08.1996 Priority country : JP

(54) WORD AND COMPOUND WORD CLASSIFYING PROCESSING METHOD, COMPOUND WORD EXTRACTING METHOD, WORD AND COMPOUND WORD CLASSIFYING PROCESSOR, SPEECH RECOGNITION SYSTEM, MACHINE TRANSLATING DEVICE, COMPOUND WORD EXTRACTING DEVICE, AND WORD AND COMPOUND WORD STORAGE MEDIUM

(57)Abstract:
PROBLEM TO BE SOLVED: To make speech recognition and machine translation accurate by classifying words and compound words included in text together and generating a class wherein the words and compound word are mixed.
SOLUTION: The word and compound word classifying processor consists of a word classifying means 1, a word class string generating means 2, a word class string extracting means 3, a token giving means 4, a word and token string generating means 5, a word and token classifying means 6, and a compound word substituting means 7. Word classes obtained by classifying words are mapped in a linear array of words of the text data to generate a linear array of word classes. In the linear array of the word classes of the text data, word class arrays which all have adherence above a specific value between adjacent word classes are extracted and tokens are given to the word class arrays. The words and tokens are classified together and then a word class array corresponding to a token is substituted by a coupla belonging to the word string. Namely, a classifying process can be performed automatically without discriminating between words and compound words.



LEGAL STATUS

[Date of request for examination] 15.07.2003
[Date of sending the examiner's decision of rejection]
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]
[Date of registration]
[Number of appeal against examiner's decision of rejection]
[Date of requesting appeal against examiner's decision of rejection]
[Date of extinction of right]

(19)【発行国】日本国特許庁(JP)
(12)【公報種別】公開特許公報(A)
(11)【公開番号】特開平10－97286
(43)【公開日】平成10年(1998)4月14日
(54)【発明の名称】単語・連語分類処理方法、連語抽出方法、単語・連語分類処理装置、音声認識装置、機械翻訳装置、連語抽出装置及び単語・連語記憶媒体
(51)【国際特許分類第6版】

G10L 3/00 561 521 G06F 17/28
【FI】

G10L 3/00 561 G 521 C G06F 15/38 Z
【審査請求】未請求
【請求項の数】17
【出願形態】OL
【全頁数】24
(21)【出願番号】特願平9－167243
(22)【出願日】平成9年(1997)6月24日
(31)【優先権主張番号】特願平8－204986
(32)【優先日】平8(1996)8月2日
(33)【優先権主張国】日本(JP)
(71)【出願人】
【識別番号】000005223
【氏名又は名称】富士通株式会社
【住所又は居所】神奈川県川崎市中原区上小田中4丁目1番1号
(72)【発明者】
【氏名】潮田 明
【住所又は居所】神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内
(74)【代理人】
【弁理士】
【氏名又は名称】大菅 義之 (外1名)

---

(57)【要約】
【課題】 単語と連語とをまとめて自動的に分類する。
【解決手段】 テキストデータにおいて出現する確率が所定値以上の単語クラス列にトークンを付与し、テキストデータの単語・トークン列に含まれる単語とトークンとが混在する集合を、テキストデータの単語・トークン列の生成確率が最大になるように分割し、トークンをテキストデータに存在する連語に置換する。

【特許請求の範囲】
【請求項1】 複数の単語の一次元列としてのテキストデータから、互いに異なるV個の単語を抽出し、前記V個の単語の集合をC個の単語クラスに分割した第1のクラスタリングを生成するステップと、前記第1のクラスタリングに基づいて生成された前記テキストデータの単語クラスの一次元列において、隣接する単語クラス間の粘着度が全て所定値以上の単語クラス列の集合を抽出するステップと、前記単語クラス列に固有のトークンを対応させ、前記単語クラス列に属する単語列を前記テキストデータから検索し、前記テキストデータの単語列を対応するトークンで置換することにより、前記テキストデータについての単語とトークンとの一次元列を生成するステップと、前記テキストデータについての単語とトークンとの一次元列において、互いに異なる単語と互いに異なるトークンとを抽出し、前記単語と前記トークンとが混在する集合を単語・トークンクラスに分割した第2のクラスタリングを生成するステップと、前記テキストデータに存在する単語列のうち、前記トークンに対応するものを連語として抽出し、前記単語・トークンクラスの中のトークンを前記連語で置換することにより、前記単語と前記連語とが混在する集合を単語・連語クラスに分割した第3のクラスタリングを生成するステップとを備えることを特徴とする単語・連語分類処理方法。
【請求項2】 前記第1のクラスタリングは、前記単語クラスの平均相互情報量に基づいて生成されることを特徴とする請求項1に記載の単語・連語分類処理方法。
【請求項3】 前記第2のクラスタリングは、前記単語・トークンクラスの平均相互情報量に基づいて生成されることを特徴とする請求項1に記載の単語・連語分類処理方法。
【請求項4】 テキストデータに含まれる単語を分類した単語クラスを生成するステップと、前記単語クラスを前記テキストデータの単語の一次元列にマッピングして単語クラスの一次元列を生成するステップと、前記テキストデータの単語クラスの一次元列において、隣接する単語クラス間の粘着度が全て所定値以上の単語クラス列を、前記テキストデータの単語クラスの一次元列から抽出するステップと、前記テキストデータに含まれる単語と前記単語クラス列とを一緒に分類するステップと、前記単語クラス列を構成する個々の単語クラスから、前記テキストデータに隣接して存在する個々の単語を別々に取り出して連語を抽出するステップと、前記単語クラス列を前記単語クラス列に属する連語で置換するステップとを備えることを特徴とする単語・連語分類処理方法。
【請求項5】 テキストデータに含まれる単語を分類した単語クラスを生成するステップと、前記単語クラスを前記テキストデータの単語の一次元列にマッピングして単語クラスの一次元列を生成するステップと、前記テキストデータの単語クラスの一次元列において、隣接する単語クラス間の粘着度が全て所定値以上の単語クラス列を、前記テキストデータの単語クラスの一次元列から抽出するステップと、前記単語クラス列を構成する個々の単語クラスから、前記テキストデータに隣接して存在する個々の単語を別々に取り出して連語を抽出するステップとを備えることを特徴とする連語抽出方法。
【請求項6】 テキストデータの単語列から互いに異なる単語を抽出し、抽出された前記単語の集合を分割して単語クラスを生成する単語分類手段と、前記テキストデータの単語の一次元列を構成する個々の単語を、前記単語が属する前記単語クラスで置換することにより、前記テキストデータの単語クラスの一次元列を生成する単語クラス列生成手段と、前記テキストデータの単語クラスの一次元列において、隣接する単語クラス間の粘着度が全て所定値以上の単語クラス列を、前記テキストデータの単語クラスの一次元列から抽出する単語クラス列抽出手段と、前記単語クラス列抽出手段により抽出された各単語クラス列にトークンを付与するトークン付与手段と、前記テキストデータの単語の一次元列のうち、前記単語クラス列抽出手段により抽出された単語クラス列に属する単語列を前記トークンで置換することにより、前記テキストデータの単語・トークンの一次元列を生成する単語・トークン列生成手段と、前記テキストデータの単語・トークンの一次元列に含まれる単語とトークンとが混在する集合を分割して単語・トークンクラスを生成する単語・トークン分類手段と、前記単語・トークンクラスの中のトークンを、前記単語・トークン列生成手段により置換された単語列に逆置換して連語を生成する連語置換手段とを備えることを特徴とする単語・連語分類処理装置。
【請求項7】 前記単語分類手段は、前記テキストデータの単語の一次元列から互いに異なる単語を抽出し、所定の出現頻度を有する単語のそれぞれに固有の単語クラスを割り当てる初期化クラス設定部と、単語クラスの集合から2つの単語クラスを取り出して仮マージする仮マージ部と、前記テキストデータの仮マージされた単語クラスについての平均相互情報量を算出する平均相互情報量算出部と、前記単語クラスの集合のうち、前記平均相互情報量が最大である2つの単語クラスを本マージする本マージ部とを備えることを特徴とする請求項6に記載の単語・連語分類処理装置。
【請求項8】 前記単語クラス列抽出手段は、前記テキストデータの単語クラスの一次元列から、隣接して存在する2つの単語クラスを順次に取り出す単語クラス取出部と、前記単語クラス取出部により取り出した2つの単語クラスの相互情報量を算出する相互情報量算出部と、前記相互情報量が所定のしきい値以上の2つの単語クラスをクラスチェーンで結ぶクラスチェーン結合部とを備えることを特徴とする請求項6に記載の単語・連語分類処理装置。
【請求項9】 前記単語・トークン分類手段は、前記テキストデータの単語・トークンの一次元列から互いに異なる単語と互いに異なるトークンとを抽出し、所定の出現頻度を有する単語とトークンとのそれぞれに固有の単語・トークンクラスを割り当てる初期化クラス設定部と、単語・トークンクラスの集合から2つの単語・トークンクラスを取り出して仮マージする仮マージ部と、前記テキストデータの仮マージされた単語・トークンクラスについての平均相互情報量を算出する平均相互情報量算出部と、前記単語・トークンクラスの集合のうち、前記平均相互情報量が最大である2つの単語・トークンクラスを本マージする本マージ部とを備えることを特徴とする請求項6に記載の単語・連語分類処理装置。
【請求項10】 テキストデータから連語を抽出する連語抽出手段と、前記テキストデータに含まれる単語と連語とを一

緒に分類して、単語と連語とが混在するクラスを生成する単語・連語分類手段とを備えることを特徴とする単語・連語分類処理装置。
【請求項11】前記クラスは、前記クラスの平均相互情報量に基づいて生成されることを特徴とする請求項10に記載の単語・連語分類処理装置。
【請求項12】テキストデータに含まれる単語を分類して単語クラスを生成する単語分類手段と、前記テキストデータの単語の一次元列を構成する個々の単語を、前記単語が属する前記単語クラスで置換することにより、前記テキストデータの単語クラスの一次元列を生成する単語クラス列生成手段と、前記テキストデータの単語クラスの一次元列において、隣接する単語クラス間の粘着度が全て所定値以上の単語クラス列を、前記テキストデータの単語クラスの一次元列から抽出する単語クラス列抽出手段と、前記単語クラス列を構成する個々の単語クラスから、前記テキストデータに隣接して存在する個々の単語を別々に取り出して連語を抽出する連語抽出手段とを備えることを特徴とする連語抽出装置。
【請求項13】前記単語クラスは、前記単語クラスの平均相互情報量に基づいて生成されることを特徴とする請求項12に記載の連語抽出装置。
【請求項14】所定のテキストデータに含まれる単語と連語とを、単語と連語とが混在するクラスに分類して格納している単語・連語辞書と、前記単語・連語辞書と所定の隠れマルコフモデルとを参照することにより、発音音声を音声認識する音声認識手段とを備えることを特徴とする音声認識装置。
【請求項15】所定のテキストデータに含まれる単語と連語とを、単語と連語とが混在するクラスに分類して格納している単語・連語辞書と、用例原文と前記用例原文に対する用例訳文とを対応させて格納している用例文集と、入力された原文の単語が属するクラスと同一のクラスに属する単語又は連語により構成される用例原文を前記用例文集から検索する用例検索手段と、前記用例原文に対する用例訳文の中の訳語を、入力された原文の単語に対する訳語に置換することにより、前記入力された原文に対する訳文を生成する用例適用手段とを備えることを特徴とする機械翻訳装置。
【請求項16】所定のテキストデータに含まれる単語と連語とを、単語と連語とが混在するクラスに分類して格納している単語・連語記憶媒体であって、前記クラスは、前記クラスの平均相互情報量に基づいて生成されていることを特徴とする単語・連語記憶媒体。
【請求項17】テキストデータの単語の一次元列から互いに異なる単語を抽出し、抽出された前記単語の集合を分割して単語クラスを生成する機能と、前記テキストデータの単語の一次元列を構成する個々の単語を、前記単語が属する前記単語クラスで置換することにより、前記テキストデータの単語クラスの一次元列を生成する機能と、前記テキストデータの単語クラスの一次元列から、隣接する単語クラス間の粘着度が全て所定値以上の単語クラス列を抽出する機能と、前記単語クラス列にトークンを付与する機能と、前記テキストデータの単語の一次元列のうち、前記単語クラス列に属する単語列を前記トークンで置換することにより、前記テキストデータの単語・トークンの一次元列を生成する機能と、前記テキストデータの単語・トークンの一次元列に含まれる単語とトークンとが混在する集合を分割して単語・トークンクラスを生成する機能と、前記単語・トークンクラスの中のトークンを、前記テキストデータに存在する単語列に逆置換して連語を生成する機能とをコンピュータに実行させるプログラムを格納したコンピュータ読み取り可能な記憶媒体。

【発明の詳細な説明】
【0001】
【発明の属する技術分野】本発明は、単語・連語分類処理方法、連語抽出方法、単語・連語分類処理装置、音声認識装置、機械翻訳装置、連語抽出装置及び単語・連語記憶媒体に関し、特に、テキストデータの中から連語を自動的に抽出し、単語及び連語を自動的に分類する場合に好適なものである。
【0002】
【従来の技術】従来の単語分類処理装置には、例えば、「Brown, P., Della Pietra, V., deSouza, P., Lai, J., Mercer, R. (1992)"Class−Based n−gram Models ofNatural Language". Computational Linguistics, Vol. 18, No4, pp. 467−479」に記載されているように、テキストデータの中で使用されている単独の単語を統計的に処理することにより、単独の単語を自動的に分類するものがあり、この単独の単語の分類結果を用いて音声認識や機械翻訳を行っていた。
【0003】
【発明が解決しようとする課題】しかしながら、従来の単語分類処理装置は、単語と連語とをまとめて自動的に分類することができず、単語と連語あるいは連語と連語の対応関係や類似度を用いて、音声認識や機械翻訳を行うことができないため、音声認識や機械翻訳を正確に実行することができないという問題があった。

【0004】そこで、本発明の第1の目的は、単語と連語とをまとめて自動的に分類することが可能な単語・連語分類処理方法及び単語・連語分類処理装置を提供することである。
【0005】また、本発明の第2の目的は、大量のテキストデータから高速に連語を抽出することが可能な連語抽出装置を提供することである。また、本発明の第3の目的は、単語と連語あるいは連語と連語の対応関係や類似度を用いることにより、正確な音声認識が可能な音声認識装置を提供することである。
【0006】また、本発明の第4の目的は、単語と連語あるいは連語と連語の対応関係や類似度を用いることにより、正確な機械翻訳が可能な機械翻訳装置を提供することである。
【0007】
【課題を解決するための手段】上述した第1の目的を達成するために、本発明によれば、テキストデータに含まれる単語と連語とを一緒に分類して、単語と連語とが混在するクラスを生成するようにしている。
【0008】このことにより、単語と単語とをまとめて分類するだけでなく、単語と連語あるいは連語と連語とをまとめて一緒に分類することができ、単語と連語あるいは連語と連語との対応関係や類似度を容易に判別することができる。
【0009】また、本発明の一態様によれば、単語を分類した単語クラスをテキストデータの単語の一次元列にマッピングして単語クラスの一次元列を生成し、テキストデータの単語クラスの一次元列において、隣接する単語クラス間の粘着度が全て所定値以上の単語クラス列を抽出してその単語クラス列にトークンを付与し、単語とトークンとを一緒に分類してから、トークンに対応する単語クラス列をその単語クラス列に属する連語で置換するようにしている。
【0010】このことにより、単語クラス列にトークンを付与してその単語クラス列を1つの単語とみなし、テキストデータに含まれる単語とトークンを付与された単語クラス列とを同等に取り扱って単語と連語との区別なく分類処理を行うことができる。また、単語を分類した単語クラスをテキストデータの単語の一次元列にマッピングして単語クラスの一次元列を生成し、隣接する単語クラス間の粘着度に基づいて連語を抽出することにより、テキストデータからの連語の抽出を高速に行うことができる。
【0011】また、上述した第2の目的を達成するために、本発明によれば、単語を分類した単語クラスをテキストデータの単語の一次元列にマッピングして単語クラスの一次元列を生成し、テキストデータの単語クラスの一次元列において、隣接する単語クラス間の粘着度が全て所定値以上の単語クラス列を抽出し、単語クラス列を構成する個々の単語クラスから、テキストデータに隣接して存在する個々の単語を別々に取り出して連語を抽出するようにしている。
【0012】このことにより、単語クラス列に基づいて連語を抽出することができ、テキストデータに存在する異なる単語の数よりも、それらの単語を分類した単語クラスの数のほうが少ないので、テキストデータの単語クラスの一次元列において、隣接する単語クラス間の粘着度が所定値以上の単語クラス列を抽出するほうが、テキストデータの単語の一次元列において、隣接する単語間の粘着度が所定値以上の単語列を抽出する場合に比べて、演算量及びメモリ容量を少なくすることができ、連語の抽出処理を高速に行うことができるとともに、メモリ資源を節約できる。なお、単語クラス列には、テキストデータの単語の一次元列に存在しない単語列が含まれている場合があるので、単語クラス列を構成する個々の単語クラスから、テキストデータに隣接して存在する個々の単語を別々に取り出して連語としている。
【0013】また、上述した第3の目的を達成するために、本発明によれば、所定のテキストデータに含まれる単語と連語とを、単語と連語とが混在するクラスに分類して格納している単語・連語辞書を参照することにより、発音音声を音声認識するようにしている。
【0014】このことにより、単語と連語あるいは連語と連語の対応関係や類似度を用いながら音声認識を行うことができ、正確な処理が可能になる。また、上述した第4の目的を達成するために、本発明によれば、所定のテキストデータに含まれる単語と連語とを、単語と連語とが混在するクラスに分類して格納している単語・連語辞書に基づいて、用例文集に格納されている用例原文と入力された原文とを対応させるようにしている。
【0015】このことにより、用例文集に格納されている用例原文の単語が連語に置き換わった原文が入力された場合においても、入力された原文に用例原文を適用して機械翻訳を行うことができ、単語と連語あるいは連語と連語の対応関係や類似度を用いた正確な機械翻訳が可能になる。
【0016】
【発明の実施の形態】以下、本発明の一実施例に係わる単語・連語分類処理装置について図面を参照しながら説明する。この実施例は、所定のテキストデータに含まれる単語と連語とを、単語と連語とが混在するクラスに分類するものである。
【0017】図1は、本発明の一実施例に係わる単語・連語分類処理装置の機能的な構成を示すブロック図である。図1において、単語分類手段1は、テキストデータの単語の一次元列から互いに異なる単語を抽出し、抽出された単語の集合を分割して単語クラスを生成する。
【0018】図2は、単語分類手段1の処理を説明するもので、テキストデータに含まれるT個の単語よりなる単語の一次元列(w1 w2 w3 w4 ・・・wT)から、テキストデータでの出現頻度順に並べたV個のボキャブラリーとしての単語{v

1 、v2 、v3 、v4 、・・・、vV ｝を生成し、このテキストデータのボキャブラリーとしての単語｛v1 、v2 、v3 、v4 、・・・、vV ｝のそれぞれに初期化クラスを割り当てる。ここで、単語の個数T個は、例えば、5000万個であり、ボキャブラリーの個数V個は、例えば、7000個である。

【0019】図2の例では、テキストデータでの出現頻度が高い、例えば、"the"、"a"、"in"、"of"が、それぞれボキャブラリーとしての単語v1 、v2、v3 、v4 に対応している。初期化クラスを割り当てられたV個のボキャブラリーとしての単語｛v1 、v2 、v3 、v4 、・・・、vV ｝は、クラスタリングによりC個の単語クラス｛C1 、C2 、C3 、C4 、・・・、CC ｝に分割される。ここで、単語クラスの個数C個は、例えば、500個である。

【0020】また、図2では、例えば、"speak"、"say"、"tell"、"talk"・・・が単語クラスC1 に分類され、"he"、"she"、"it"・・・が単語クラスC5 に分類され、"car"、"track"、"wagon"・・・が単語クラスC32に分類され、"Toyota"、"Nissan"、"GM"・・・が単語クラスC300 に分類されている例を示している。

【0021】このV個のボキャブラリーとしての単語｛v1 、v2 、v3 、v4 、・・・、vV ｝よりなる単語の分類は、例えば、テキストデータに存在する2つの単語がおのおの属する2つの単語クラスをマージした場合、元のテキストデータの生成確率の減少が最も少なくなるものを同一の単語クラスに統合することにより行う。ここで、元のテキストデータのクラスバイモデルによる生成確率は、平均相互情報量AMIを用いて表現することができ、この平均相互情報量AMIは以下の式により表すことができる。

【0022】

【数1】

【0023】ここで、Pr(Ci)は、テキストデータの単語の一次元列(w1 w2 w3 w4 ・・・wT)をその単語が属する単語クラスで置き換えた場合、そのテキストデータの単語クラスの一次元列でのクラスCi の出現確率、Pr(Cj)は、テキストデータの単語の一次元列(w1 w2 w3 w4 ・・・wT)をその単語が属する単語クラスで置き換えた場合、そのテキストデータの単語クラスの一次元列でのクラスCj の出現確率、Pr(Ci 、Cj)は、テキストデータの単語の一次元列(w1 w2 w3 w4・・・wT)を、その単語が属する単語クラスで置き換えた場合、そのテキストデータの単語クラスの一次元列での単語クラスCi の次に隣接して単語クラスCj が出現する確率である。

【0024】図3は、図1の単語分類手段1の機能的な構成の一例を示すブロック図である。図3において、初期化クラス設定部10は、テキストデータの単語の一次元列｛w1 w2 w3 w4 ・・・wT ｝から互いに異なる単語を抽出し、所定の出現頻度を有する単語｛v1 、v2 、v3 、v4 、・・・、vV ｝のそれぞれに固有の単語クラス｛C1 、C2 、C3 、C4 、・・・、CV ｝を割り当てる。

【0025】仮マージ部11は、単語クラスの集合｛C1 、C2 、C3 、C4 、・・・、CM ｝から2つの単語クラス｛Ci 、Cj ｝を取り出して仮マージする。平均相互情報量算出部12は、テキストデータの仮マージされた単語クラス｛C1 、C2 、C3 、C4 、・・・、CM-1 ｝についての平均相互情報量AMIを(1)式により算出する。この場合、M個の単語クラスの集合｛C1 、C2 、C3 、C4 、・・・、CM ｝から2つの単語クラス｛Ci 、Cj ｝を取り出だす取り出しかたは、M(M－1)／2個だけ存在するので、M(M－1)／2回の平均相互情報量AMIの計算を行う必要がある。

【0026】本マージ部13は、仮マージにより計算されたM(M－1)／2個の平均相互情報量AMIの基づいて、平均相互情報量AMIを最大とする2つの単語クラス｛Ci 、Cj ｝を単語クラスの集合｛C1 、C2 、C3 、C4 、・・・、CM ｝から取り出して本マージする。このことにより、本マージされたいずれかの単語クラス｛Ci 、Cj ｝に属する単語は、同一の単語クラスに分類される。

【0027】図1の単語クラス列生成手段2は、テキストデータの単語列(w1 w2 w3 w4 ・・・wT)を構成する個々の単語を、単語が属する単語クラス｛C1 、C2、C3 、C4 、・・・、CV ｝で置換することにより、テキストデータの単語クラス列を生成する。

【0028】図4は、テキストデータの単語クラスの一次元列の一例を示す図である。図4において、単語分類手段1によりC個の単語クラス｛C1 、C2 、C3 、C4 、・・・、CC ｝が生成されているものとし、例えば、単語クラスC1 には、ボキャブラリーv1 、v37、・・・が属しており、単語クラスC2 には、ボキャブラリーv3 、v15、・・・が属しており、単語クラスC3 には、ボキャブラリーv2 、v4 、・・・が属しており、単語クラスC4 には、ボキャブラリーv7 、v9 、・・・が属しており、単語クラスC5 には、ボキャブラリーv6 、v8 、v26、vV 、・・・が属しており、単語クラスC6 には、ボキャブラリーv6 、v23、・・・が属しており、単語クラスC7 には、ボキャブラリーv5 、v10、・・・が属しているものとする。

【0029】また、テキストデータの単語の一次元列(w1 w2 w3 w4 ・・・wT)において、例えば、単語w1 が示すボキャブラリーとしての単語がv15、単語w2 が示すボキャブラリーとしての単語がv2 、単語w3 が示すボキャブラリーとしての単語がv23、単語w4 が示すボキャブラリーとしての単語がv4 、単語w5 が示すボキャブラリーとしての単語がv5 、単語w6 が示すボキャブラリーとしての単語がv15、単語w7 が示すボキャブラリーとしての単語がv5 、単語w8 が示すボキャブラリーとしての単語がv26、単語w9 が示すボキャブラリーとしての単語がv37、単語w10 が示すボキャブラリーとしての単語がv2 、・・・、単語wT が示すボキャブラリーとしての単語がv8 であるとする。

【0030】この場合、ボキャブラリーv15 は単語クラスC2 に属しているので、単語w1 は単語クラスC2 にマッピングされ、ボキャブラリーv2 は単語クラスC3 に属しているので、単語w2 は単語クラスC3 にマッピングされ、ボキャブラリーv23 は単語クラスC6 に属しているので、単語w3 は単語クラスC6 にマッピングされ、ボキャブラリーv4 は単語クラスC3 に属しているので、単語w4 は単語クラスC3 にマッピングされ、ボキャブラリーv5 は単語クラスC7 に属しているので、単語w5 は単語クラスC7 にマッピングされ、ボキャブラリーv15 は単語クラスC2 に属しているので、単語w6 は単語クラスC2 にマッピングされ、ボキャブラリーv5 は単語クラスC7 に属しているので、単語w7 は単語クラスC7 にマッピングされ、ボキャブラリーv26 は単語クラスC5 に属しているので、単語w8 は単語クラスC5 にマッピングされ、ボキャブラリーv37 は単語クラスC1 に属しているので、単語w9 は単語クラスC1 にマッピングされ、ボキャブラリーv2 は単語クラスC3 に属しているので、単語w10 は単語クラスC3 にマッピングされ、・・・、ボキャブラリーv8 は単語クラスC5 に属しているので、単語wT は単語クラスC5 にマッピングされる。

【0031】すなわち、テキストデータの単語の一次元列(w1 w2 w3 w4 ・・・wT)が、C個の単語クラス｛C1 、C2 、C3 、C4 、・・・、CC ｝によりマッピングされた結果として、テキストデータの単語クラスの一次元列(C2 C3 C6C3 C7 C2 C7 C5 C1 C3 ・・・C5)が1対1対応で生成される。

【0032】図1の単語クラス列抽出手段3は、テキストデータの単語クラスの一次元列においての単語クラス間の粘着度が全て所定値以上の単語クラス列を、テキストデータの単語クラスの一次元列から抽出する。ここで、単語クラス間の粘着度は、単語クラス列を構成する単語クラス間のつながりの強さを示す指標であり、この粘着度を表現するものとして、例えば、相互情報量MI、相関係数、コサインメジャー、liklihood ratioなどがある。

【0033】以下の説明では、単語クラス間の粘着度として、相互情報量MIを用いることにより、テキストデータの単語クラスの一次元列から単語クラス列を抽出する場合を例にとる。

【0034】図5は、単語クラス列抽出手段3により抽出された単語クラス列の一例を示す図である。図5において、テキストデータの単語の一次元列(w1 w2 w3 w4 w5 w6w7 ・・・wT)に対してマッピングされた結果として、テキストデータの単語クラスの一次元列(C2 C3 C6 C3 C7 C2 C7 ・・・C5)が1対1対応で生成されているものとする。このテキストデータの単語クラスの一次元列(C2C3 C6 C3 C7 C2 C7 ・・・C5)から、隣接する2つの単語クラス(Ci、Cj)を順次に取り出し、隣接する2つの単語クラス(Ci 、Cj)についての相互情報量MI(Ci 、Cj)を、以下の(2)式により計算する。

【0035】

MI(Ci 、Cj)

＝log｛Pr(Ci 、Cj)／(Pr(Ci)Pr(Cj))｝

・・・(2)

そして、隣接する2つの単語クラス(Ci 、Cj)についての相互情報量MI(Ci 、Cj)が所定のしきい値TH以上の場合、これら隣接する2つの単語クラス(Ci 、Cj)をクラスチェーンで結んで互いに関連づける。

【0036】例えば、図5において、隣接する2つの単語クラス(C2 、C3)についての相互情報量MI(C2 、C3)、隣接する2つの単語クラス(C3 、C6)についての相互情報量MI(C3 、C6)、隣接する2つの単語クラス(C6 、C3)についての相互情報量MI(C6 、C3)、隣接する2つの単語クラス(C3、C7)についての相互情報量MI(C3 、C7)、隣接する2つの単語クラス(C7 、C2)についての相互情報量MI(C7 、C2)、隣接する2つの単語クラス(C2 、C7)についての相互情報量MI(C2 、C7)、・・・を(2)式により順次に計算する。

【0037】そして、相互情報量MI(C2 、C3)、相互情報量MI(C3 、C7)、相互情報量MI(C7 、C2)、・・・がしきい値TH以上で、相互情報量MI(C3 、C6)、相互情報量MI(C6 、C3)、相互情報量MI(C2 、C7)、・・・がしきい値THより小さい場合、隣接する2つの単語クラス(C2 、C3)、(C3 、C7)、(C7 、C2)、・・・をそれぞれクラスチェーンで結ぶことにより、単語クラス列C2 －C3 、C3 －C7 －C2 、・・・を抽出する。

【0038】図6は、図1の単語クラス列抽出手段3の機能的な構成の一例を示すブロック図である。図6において、単語クラス取出部30は、テキストデータの単語クラスの一次元列から、隣接して存在する2つの単語クラス(Ci 、Cj)を順次に取り出す。

【0039】相互情報量算出部31は、単語クラス取出部30により取り出した2つの単語クラス(Ci 、Cj)の相互情報量MI(Ci 、Cj)を(2)式により算出する。

【0040】クラスチェーン結合部32は、相互情報量MI(Ci 、Cj)が所定のしきい値以上の2つの単語クラス(Ci 、Cj)をクラスチェーンで結ぶ。図1のトークン付与手段4は、単語クラス列抽出手段3によりクラスチェーンで結ばれた単語クラス列にトークンを付与する。

【0041】図7は、トークン付与手段4により付与されたトークンの一例を示す図である。図7において、クラスチェーン

で結ばれた単語クラス列は、例えば、C1 −C3 、C1 −C7 、・・・、C2 −C3 、C2 −C11、・・・、C300 −C32、・・・、C1 −C3 −C80、C1 −C4 −C5 、C3 −C7 −C2 、・・・、C1−C9 −C11−C32、・・・とする。この場合、単語クラス列C1 −C3 に対してトークンt1 を付与し、単語クラス列C1 −C7 に対してトークンt2 を付与し、・・・、単語クラス列C2 −C3 に対してトークンt3 を付与し、単語クラス列C2 −C11に対してトークンt4 を付与し、・・・、単語クラス列C300 −C32 に対してトークンt5 を付与し、、・・・、単語クラス列C1 −C3 −C80 に対してトークンt6 を付与し、単語クラス列C1 −C4 −C5 に対してトークンt7 を付与し、単語クラス列C3 −C7 −C2 に対してトークンt8 を付与し、・・・、単語クラス列C1 −C9 −C11−C32 に対してトークンt9 を付与する。

【0042】図1の単語・トークン列生成手段5は、テキストデータの単語の一次元列(w1 w2 w3 w4 w5 w6 w7 ・・・ wT)のうち、単語クラス列抽出手段4により抽出された単語クラス列に属する単語列をトークンで置換することにより、テキストデータの単語・トークンの一次元列を生成する。

【0043】図8は、テキストデータの単語・トークンの一次元列の一例を示す図である。図8において、テキストデータの単語の一次元列(w1 w2 w3 w4 w5 w6w7 ・・・wT)に対してマッピングされた結果として、テキストデータの単語クラスの一次元列(C2 C3 C6 C3 C7 C2 C7 ・・・C5)が1対1対応で生成されているものとし、クラスチェーンで結ばれた単語クラス列C2 −C3 、C3 −C7 −C2 、・・・に対して、図7に示すように、トークンt3 、t8 、・・・が付与されているものとする。

【0044】この場合、クラスチェーンで結ばれた単語クラス列C2 −C3 に属するテキストデータの単語列(w1 w2)をトークンt3 で置き換え、クラスチェーンで結ばれた単語クラス列C3 −C7 −C2 に属するテキストデータの単語列(w4 w5 w6)をトークンt8 で置き換えることにより、テキストデータの単語・トークンの一次元列(t3 w3 t8 w7 ・・・ wT)を生成する。

【0045】図9は、テキストデータの単語・トークンの一次元列の一例を英文を例にとって示す図である。図9(b)のテキストデータの単語の一次元列(w1 w2 w3 w4 w5 w6 w7 w8 w9 w10w11w12w13w14w15)として、図9(a)の"He wentto the apartment by bus and she went to New York by plane"が対応しているものとし、この単語の一次元列(w1 w2 w3 w4 w5 w6 w7 w8 w9 w10w11w12w13w14w15)に1対1で対応する単語クラスの一次元列が図9(c)の(C5 C90C3 C21C18C101 C32C2 C5 C90C3 C63C28C101 C32)で与えられるものとする。

【0046】この単語クラスの一次元列(C5 C90C3 C21C18C101 C32C2 C5 C90C3 C63C28C101 C32)において、隣接する2つの単語クラス(Ci 、Cj)の相互情報量MI(Ci 、Cj)を計算し、相互情報量MI(C63、C28)が所定のしきい値TH以上、相互情報量MI(C5 、C90)、MI(C90、C3)、MI(C3 、C21)、MI(C21、C18)、MI(C18、C101)、MI(C101、C32)、MI(C32、C2)、MI(C2 、C5)、MI(C5 、C90)、MI(C90、C3)、MI(C3 、C63)、MI(C28、C101)及びMI(C101、C32)が所定のしきい値THより小さい場合、隣接する2つの単語クラス(C63、C28)が、図9(d)に示すように、クラスチェーンで結ばれる。

【0047】このクラスチェーンで結ばれた2つの単語クラス(C63、C28)はトークンt1 に置き換えられ、図9(e)に示すように、単語・トークンの一次元列(w1w2 w3 w4 w5 w6 w7 w8 w9 w10w11t1 w14w15)が生成される。

【0048】図1の単語・トークン分類手段6は、テキストデータの単語・トークンの一次元列のN個の単語の集合{w1 、w2 、w3 、w4 、・・・、wN }又はL個のトークンの集合{t1 、t2 、t3 、t4 、・・・、tL }を分割することにより、単語とトークンとが混在して存在するD個の単語・トークンクラス{T1 、T2 、T3 、T4 、・・・、TD }を生成する。

【0049】この単語・トークン分類手段6では、トークンを付与された単語クラス列が1つの単語のようにみなされ、テキストデータに含まれる単語{w1 、w2 、w3 、w4 、・・・、wN }とトークン{t1 、t2 、t3 、t4 、・・・、tL }とを同等に取り扱うことができるので、単語{w1 、w2 、w3 、w4 、・・・、wN }とトークン{t1 、t2 、t3 、t4 、・・・、tL }との区別なく分類処理を行うことができる図10は、図1の単語・トークン分類手段6の機能的な構成を示すブロック図である。

【0050】図10において、初期化クラス設定部40は、テキストデータの単語・トークン列から互いに異なる単語と互いに異なるトークンとを抽出し、所定の出現頻度を有するN個の単語{w1 、w2 、w3 、w4 、・・・、wN }とL個のトークン{t1 、t2 、t3 、t4 、・・・、tL }とのそれぞれに固有の単語・トークンクラス{T1 、T2 、T3 、T4 、・・・、TY }を割り当てる。

【0051】仮マージ部41は、単語・トークンクラスの集合{T1 、T2 、T3 、T4 、・・・、TM }から2つの単語・トークンクラス{Ti 、Tj }を取り出して仮マージする。

【0052】平均相互情報量算出部42は、テキストデータの仮マージされた単語・トークンクラス{T1 、T2 、T3 、T4 、・・・、TM-1 }についての平均相互情報量AMIを(1)式により算出する。この場合、M個の単語クラス・トークンクラスの集合{T1 、T2 、T3 、T4 、・・・、TM }から、2つの単語・トークンクラス{Ti 、Tj }を取り出だす取り出しかたは、M(M−1)／2個だけ存在するので、M(M−1)／2回の平均相互情報量AMIの計算を行う必要がある。

【0053】本マージ部43は、仮マージにより計算されたM(M−1)／2個の平均相互情報量AMIの基づいて、平均相互情報量AMIを最大とする2つの単語・トークンクラス{Ti 、Tj }を単語クラス・トークンクラスの集合{T1 、T2 、T3 、T4 、・・・、TM }から取り出して本マージする。このことにより、本マージされたいずれかの単語・トークンクラス{Ti 、Tj }に属する単語及びトークンは、同一の単語クラス・トークンクラスに分類される。

【0054】図1の連語置換手段7は、単語・トークンクラスの中のトークンを、単語・トークン列生成手段5により置換された単語列に逆置換して連語を生成する。図11は、クラスチェーンと連語との関係を説明する図である。

【0055】図11において、例えば、単語クラスC300 と単語クラスC32 とがクラスチェーンで結ばれ、このクラスチェーンで結ばれた単語クラス列C300 −C32 にトークンt5 が付与されているとする。また、単語"Toyota"、"Nissan"、"GM"・・・などのA個の単語が単語クラスC300 に属し、単語"car"、"track"、"wagon"・・・などのB個の単語が単語クラスC32 に属しているものとする。

【0056】この場合、連語の候補として、図11(b)に示すように、"Toyotacar"、"Toyota track"、"Toyota wagon"、"Nissan car"、"Nissan track"、"Nissanwagon"、"GM car"、"GM track"、"GM wagon"、・・・など、単語クラスC300 に属するA個の単語と単語クラスC32に属するB個の単語との順列の数A×Bだけ連語の候補が生成される。この連語の候補の中にはテキストデータに存在しない連語も含まれているので、テキストデータをスキャンすることにより、これらの連語の候補からテキストデータに存在する連語のみを抽出する。例えば、テキストデータには、"Nissan track"及び"Toyota wagon"は存在するが、"Toyota car"、"Toyota track"、"Nissan car"、"Nissan wagon"、"GM car"、"GM track"及び"GM wagon"は存在しない場合、図11(c)に示すように、"Nissan track"及び"Toyota wagon"のみが連語としてテキストデータから抽出される。

【0057】図12は、C個の単語クラス{C1 、C2 、C3 、C4 、・・・、CC }、D個の単語・トークンクラス{T1 、T2 、T3 、T4 、・・・、TD }及びD個の単語・連語クラス{R1 、R2 、R3 、R4 、・・・、RD }の一例を示す図である。

【0058】図12(a)において、C個の単語クラス{C1 、C2 、C3 、C4 、・・・、CC }が、図1の単語分類手段1により生成され、例えば、"he"、"she"、"it"・・・などの単語が単語クラスC5 に属し、"York"、"London"・・・などの単語が単語クラスC28 に属し、"car"、"track"、"wagon"・・・などの単語が単語クラスC32 に属し、"new"、"old"・・・などの単語が単語クラスC63 に属し、"Toyota"、"Nissan"、"GM"・・・などの単語が単語クラスC300 に属しているものとする。また、テキストデータには、"New York"、"Nissantrack"及び"Toyota wagon"の連語が多数存在しているものとする。

【0059】このC個の単語クラス{C1 、C2 、C3 、C4 、・・・、CC }をテキストデータの単語の一次元列(w1 w2 w3 w4 ・・・wT)に1対1対応でマッピングした単語クラスの一次元列において、図1の単語クラス列抽出手段3は、"new"が属する単語クラスC63 と"York"が属する単語クラスC28 との粘着度が大きいと判断し、単語クラスC63 と単語クラスC28 とをクラスチェーンで結ぶ。また、単語クラス列抽出手段3は、"Toyota"及び"Nissan"が属する単語クラスC300 と"track"及び"wagon"が属する単語クラスC32 との粘着度が大きいと判断し、単語クラスC300 と単語クラスC32 とをクラスチェーンで結ぶ。

【0060】トークン付与手段4は、単語クラス列C63−C28 にトークンt1 を付与し、単語クラス列C300 −C32 にトークンt5 を付与する。単語・トークン列生成手段5は、テキストデータの単語の一次元列(w1 w2w3 w4 ・・・wT)に存在する"New York"をトークンt1 で置き換え、テキストデータの単語の一次元列(w1 w2 w3 w4 ・・・wT)に存在する"Nissan track"及び"Toyota wagon"をトークンt5 で置き換えた単語・トークンの一次元列を生成する。

【0061】単語・トークン分類手段6は、この単語・トークンの一次元列に存在する"he"、"she"、"it"、"London"、"car"、"track"、"wagon"・・・などの単語及び"t1 "、"t5 "などのトークンについての分類処理を行い、図12(b)のD個の単語・トークンクラス{T1 、T2 、T3 、T4 、・・・、TD }を生成する。

【0062】単語・トークンクラス{T1 、T2 、T3 、T4 、・・・、TD }において、例えば、"he"、"she"、"it"・・・などの単語やトークンが単語・トークンクラスT5 に属し、"t1 "、"London"・・・などの単語やトークンが単語・トークンクラスT28 に属し、"car"、"track"、"wagon"、"t5 "・・・などの単語やトークンが単語・トークンクラスT32 に属し、"new"、"old"・・・などの単語やトークンが単語・トークンクラスT63 に属し、"Toyota"、"Nissan"、"GM"・・・などの単語やトークンが単語・トークンクラスT300 に属している。このように、単語・トークンクラス{T1 、T2 、T3 、T4 、・・・、TD }には、単語とトークンとの区別なく、単語とトークンとが混在して分類されている。

【0063】連語置換手段7は、図12(b)の単語・トークンクラス{T1 、T2 、T3、T4 、・・・、TD }に存在する"t1 "、"t5 "などのトークンを、テキストデータの単語の一次元列に存在する連語で逆置換することにより、図12(c)の単語・連語クラス{R1 、R2 、R3 、R4 、・・・、RD }を生成する。例えば、単語・トークンクラスT28 に属しているトークンt1 は、単語・トークン列生成手段5により、テキストデータの単語の一次元列に存在する"New York"と置換されたものなので、このトークンt1 を"New York"で逆置換することにより、単語・連語クラスR28 を生成し、単語・トークンクラスT32 に属しているトークンt5 は、単語・トークン列生成手段5により、テキストデータの単語の一次元列に存在す

る"Nissan track"及び"Toyota wagon"と置換されたものなので、このトークンt5 を"Nissan track"及び"Toyot a wagon"で逆置換することにより、単語・連語クラスR32 を生成する。

【0064】図13は、図1の単語・連語分類処理装置を実現するシステム構成を示すブロック図である。図13において、単語・連語分類処理部41のメモリインターフェース42、46、CPU43、ROM44、ワークRAM45、RAM47、ドライバ71及び通信インタフェース72はバス48を介して互いに接続され、テキストデータ40が単語・連語分類処理部41に入力されると、ROM44に格納されているプログラムに従って、CPU43はテキストデータ40を処理し、テキストデータ40の単語及び連語の分類処理を行う。テキストデータ40の単語及び連語の分類処理結果は、単語・連語辞書49に格納される。なお、テキストデータ40や単語及び連語の分類処理結果を通信インタフェース72から通信ネットワーク73を介して送信したり、受信したりすることも可能である。

【0065】また、単語及び連語の分類処理を行うプログラムを、ハードディスク74、ICメモリカード75、磁気テープ76、フロッピーディスク77またはCD－ROMやDVD－ROMなどの光ディスク78による記憶媒体からRAM47にロードした後、このプログラムをCPU43で実行させるようにしてもよい。

【0066】さらに、単語及び連語の分類処理を行うプログラムを、通信インタフェース72を介して通信ネットワーク73から取り出すこともできる。通信インタフェース72と接続される通信ネットワーク73として、例えば、LAN(LocalArea Network)、WAN(Wide Area Network)、インターネット、アナログ電話網、デジタル電話網(ISDN:Integral Service Digital Network)、PHS(パーソナルハンディシステム)や衛星通信などの無線通信網などを用いることが可能である。

【0067】図14は、図1の単語・連語分類処理装置の動作を示すフローチャートである。図14において、まず、ステップS1に示すように、単語クラスタリング処理を行う。この単語クラスタリング処理では、複数の単語の一次元列(w1 w2 w3 w4 ･･･wT)としてのテキストデータから、互いに異なるV個の単語{v1 、v2 、v3 、v4 、･･･、vV }を抽出し、V個の単語の集合{v1 、v2 、v3 、v4 、･･･、vV }をC個の単語クラス{C1 、C2 、C3 、C4、･･･、CC }に分割する第1のクラスタリング処理を行う。

【0068】ここで、V個の単語{v1 、v2 、v3 、v4 、･･･、vV }それぞれに単語クラス{C1 、C2 、C3 、C4 、･･･、CV }を割り当ててから、V個の単語クラス{C1 、C2 、C3 、C4 、･･･、CV }についてマージ処理を行うことにより、V個の単語クラス{C1 、C2 、C3 、C4 、･･･、CV }の個数を1つずつ減らしてC個の単語クラス{C1 、C2 、C3 、C4 、･･･、CC }を生成する場合、Vが7000もの数となって大きなものとなるときは、マージ処理を行うための(1)式の平均相互情報量AMIの計算回数が莫大なものとなり、現実的ではなくなる。このため、ウィンドウ処理を行って、マージ処理を行う単語クラスの数を減らすようにする。

【0069】図15は、ウィンドウ処理を説明する図である。図15(a)において、テキストデータのV個の単語{v1 、v2 、v3 、v4 、･･･、vV }それぞれに割り当てられたV個の単語クラス{C1 、C2 、C3 、C4 、･･･、CV }のうち、テキストデータでの出現頻度の大きい単語に割り当てられたC＋1個の単語クラス{C1 、C2 、C3 、C4 、･･･、CC 、CC+1 }を取り出し、このC＋1個の単語クラス{C1 、C2 、C3 、C4、･･･、CC 、CC+1 }についてのマージ処理を行う。

【0070】ここで、図15(b)に示すように、M個の単語クラス{C1 、C2 、C3 、C4 、･･･、CM }は、ウィンドウ内のC＋1個の単語クラス{C1 、C2 、C3 、C4 、･･･、CC 、CC+1 }についてのマージ処理を行った場合、M個の単語クラス{C1 、C2 、C3 、C4 、･･･、CM }の数が1つ減ってM－1個の単語クラス{C1 、C2 、C3 、C4 、･･･、CM-1 }となるとともに、ウィンドウ内のC＋1個の単語クラス{C1 、C2 、C3 、C4 、･･･、CC 、CC+1 }の数も1つ減ってC個の単語クラス{C1 、C2 、C3 、C4 、･･･、CC }となる。

【0071】この場合、図15(c)に示すように、ウィンドウ外の単語クラス{CC+1 、･･･、CM-1 }のうち、テキストデータでの出現頻度が最も大きい単語クラスCC+1 をウィンドウ内に入れ、ウィンドウ内の単語クラスの数が一定に保たれるようにする。

【0072】そして、ウィンドウ外に単語クラスがなくなり、図15(d)のC個の単語クラス{C1 、C2 、C3 、C4 、･･･、CC }が生成された時に、単語クラスタリング処理を終了する。

【0073】なお、上述した実施例では、ウィンドウ内の単語クラスの個数をC＋1個に設定したが、C＋1個以外のV個未満の数でもよく、また、途中で変化させるようにしてもよい。

【0074】図16は、ステップS1の単語クラスタリング処理を示すフローチャートである。図16において、まず、ステップS10に示すように、T個の単語の一次元列(w1 w2 w3 w4 ･･･wT)としてのテキストデータに基づいて、重複を除いた全てのV個の単語{v1 、v2 、v3 、v4 、･･･、vV }の出現頻度を調べ、これらのV個の単語{v1 、v2 、v3 、v4 、･･･、vV }を出現頻度の高い単語から順に並べて、これらのV個の単語{v1 、v2 、v3 、v4 、･･･、vV }のそれぞれをV個の単語クラス{C1 、C2 、C3 、C4 、･･･、CV }に割り当てる。

【0075】次に、ステップS11に示すように、V個の単語クラス{C1 、C2 、C3 、C4 、･･･、CV }の単語のうち、出現頻度の高い単語クラスの単語から、V個未満のC＋1個の単語クラスの単語を1つのウィンドウ内の単語クラスの単語とする。

【0076】次に、ステップS12に示すように、1つのウィンドウ内の単語クラスの単語の中で、全ての組み合わせの仮ペアを作り、各仮ペアを仮マージした時の平均相互情報量AMIを(1)式により計算する。

【0077】次に、ステップS13に示すように、全ての組み合わせの仮ペアについての平均相互情報量AMIのうち、最大となる平均相互情報量AMIを有する仮ペアを本マージすることにより、単語クラスを1つだけ減らし、本マージ後の1つのウィンドウ内の単語クラスの単語を更新する。

【0078】次に、ステップS14に示すように、ウィンドウ外の単語クラスはなくなり、かつ、ウィンドウ内の単語クラスはC個になったかどうかを判断し、この条件が成り立たない場合、ステップS15に進み、現在のウィンドウよりも外側にあり、最大の出現頻度を有するクラスの単語をウィンドウ内に入れ、ステップS12に戻り、以上の処理を繰り返すことにより、単語クラスの数を減少させる。

【0079】一方、ステップS14の条件が成り立ち、ウィンドウ外に単語クラスがなくなり、単語クラスの数がC個となった場合、ステップS16に進み、ウィンドウ内のC個の単語クラス{C1 、C2 、C3 、C4 、･･･、CC }をメモリに記憶する。

【0080】次に、図14のステップS2に示すように、クラスチェーン抽出処理を行う。このクラスチェーン抽出処理では、ステップS1の第1のクラスタリング処理に基づいて生成されたテキストデータの単語クラスの一次元列において、所定のしきい値以上の相互情報量を有する隣接する2つの単語クラスをチェーンで結ぶことにより、チェーンで結ばれた単語クラス列の集合を抽出する。

【0081】図17は、ステップS2のクラスチェーン抽出処理の第1実施例を示すフローチャートである。図17において、まず、ステップS20に示すように、テキストデータの単語クラスの一次元列から、互いに隣接する2つの単語クラス(C i 、Cj)を取り出す。

【0082】次に、ステップS21に示すように、ステップS20で取り出した2つの単語クラス(Ci 、Cj)についての相互情報量MI(Ci 、Cj)を(2)式により計算する。

【0083】次に、ステップS22に示すように、ステップS21で計算した相互情報量MI(Ci 、Cj)が所定のしきい値TH以上であるかどうかを判断し、相互情報量MI(Ci 、Cj)が所定のしきい値TH以上である場合、ステップS23に進んで、ステップS20で取り出した2つの単語クラス(Ci 、Cj)をクラスチェーンで結んでメモリに格納し、相互情報量MI(Ci 、Cj)が所定のしきい値THより小さい場合、ステップS23をスキップする。

【0084】次に、ステップS24に示すように、メモリに格納されているクラスチェーンで結ばれた単語クラスにおいて、単語クラスCi で終了しているクラスチェーンが存在するかどうかを判断し、単語クラスCi で終了しているクラスチェーンが存在する場合、ステップS25に進んで、単語クラスCi で終了しているクラスチェーンに単語クラスCj をつなぐ。

【0085】一方、ステップS24において、単語クラスCi で終了しているクラスチェーンが存在しない場合、ステップS25をスキップする。次に、ステップS26に示すように、テキストデータの単語クラスの一次元列から、互いに隣接する2つの単語クラス(Ci 、Cj)を全て取り出したかどうかを判断し、互いに隣接する2つの単語クラス(Ci 、Cj)を全て取り出した場合、クラスチェーン抽出処理を終了し、互いに隣接する2つの単語クラス(Ci 、Cj)を全て取り出していない場合、ステップS20に戻って以上の処理を繰り返す。

【0086】図18は、ステップS2のクラスチェーン抽出処理の第2実施例を示すフローチャートである。図18において、まず、ステップS201に示すように、テキストデータの単語クラスの一次元列から、互いに隣接する2つの単語クラス(Ci 、Cj)を順次に取り出す。そして、取り出した2つの単語クラス(Ci 、Cj)について、相互情報量MI(Ci 、Cj)を(2)式により計算することにより、長さ2の全てのクラスチェーンをテキストデータの単語クラスの一次元列から抽出する。

【0087】次に、ステップS202に示すように、長さ2の全てのクラスチェーンをそれぞれオブジェクトで置き換える。ここで、オブジェクトは、上述したトークンと同じものを表しているが、長さ2のクラスチェーンに付与されたトークンを、特に、オブジェクトと呼ぶ。

【0088】次に、ステップS203に示すように、テキストデータのクラスの一次元列に対し、ステップS202でオブジェクトが付与された長さ2のクラスチェーンをオブジェクトで置き換え、テキストデータのクラスとオブジェクトの一次元列を生成する。

【0089】次に、ステップS204に示すように、テキストデータのクラスとオブジェクトの一次元列の中に存在する1つのオブジェクトを1つのクラスとみなし、2つのクラス(Ci 、Cj)についての相互情報量MI(Ci 、Cj)を(2)式により計算する。すなわち、テキストデータのクラスとオブジェクトの一次元列においての相互情報量MI(Ci 、Cj)は、互いに隣接する1つのクラスと1つのクラスとの間で算出される場合、互いに隣接する1つのクラスと1つのオブジェクト(長さ2のクラスチェーン)との間で算出される場合、及び互いに隣接する1つのオブジェクト(長さ2のクラスチェーン)と1つ

のオブジェクト(長さ2のクラスチェーン)との間で算出される場合がある。
【0090】次に、ステップS205に示すように、ステップS204で計算した相互情報量MI(Ci 、Cj)が所定のしきい値TH以上であるかどうかを判断し、相互情報量MI(Ci 、Cj)が所定のしきい値TH以上である場合、ステップS26に進んで、ステップS204で取り出した互いに隣接する2つのクラス、又は互いに隣接する1つのクラスと1つのオブジェクト、又は互いに隣接する2つのオブジェクトをクラスチェーンで結び、相互情報量MI(Ci 、Cj)が所定のしきい値THより小さい場合、ステップS206をスキップする。
【0091】図19は、テキストデータのクラスとオブジェクトの一次元列において抽出されたクラスチェーンを示す図である。図19において、互いに隣接する1つのクラスと1つのクラスとの間でクラスチェーンが抽出された場合、長さ2のクラスチェーン(オブジェクト)が生成され、互いに隣接する1つのクラスと1つのオブジェクトとの間でクラスチェーンが抽出された場合、長さ3のクラスチェーンが生成され、互いに隣接する1つのオブジェクトと1つのオブジェクトとの間でクラスチェーンが抽出された場合、長さ4のクラスチェーンが生成される。

【0092】次に、図18のステップS207に示すように、クラスチェーン抽出処理が所定の回数行われたかどうかを判断し、所定の回数行われていない場合は、ステップS202に戻って以上の処理を繰り返す。
【0093】このように、長さ2のクラスチェーンをオブジェクトに置き換えて、相互情報量MI(Ci 、Cj)を算出することを繰り返すことにより、任意の長さのクラスチェーンを抽出することができる。
【0094】次に、図14のステップS3に示すように、トークン置換処理を行う。このトークン置換処理では、ステップS2のクラスチェーン抽出処理で抽出された単語クラス列に固有のトークンを対応させ、この単語クラス列に属する単語列をテキストデータの単語の一次元列から検索し、テキストデータの単語列を対応するトークンで置換することにより、テキストデータについての単語とトークンとの一次元列を生成する。
【0095】図20は、ステップS3のトークン置換処理を示すフローチャートである。図20において、まず、ステップS30に示すように、抽出されたクラスチェーンを重複を除いて所定の規則でソートし、それぞれのクラスチェーンにトークンを対応させて、クラスチェーンに名前を付ける。ここで、クラスチェーンのソートは、例えば、ASCIIコード順で行う。
【0096】次に、ステップS31に示すように、トークンに対応させたクラスチェーンを1つ取り出す。次に、ステップS32に示すように、テキストデータの単語の一次元列の中にクラスチェーンで結ばれた単語クラス列に属する単語列が存在するかどうかを判断し、クラスチェーンで結ばれた単語クラス列に属する単語列が存在する場合、ステップS33に進み、テキストデータの対応する単語列を1つのトークンで置き換え、クラスチェーンで結ばれた単語クラス列に属する単語列がテキストデータの単語の一次元列の中に存在しなくなるまで以上の処理を繰り返す。
【0097】一方、クラスチェーンで結ばれた単語クラス列に属する単語列が存在しない場合、ステップS34に進み、ステップS30でトークンに対応させた全てのクラスチェーンについての連語・トークン置換処理が終了したかどうかを判断し、全てのクラスチェーンについての連語・トークン置換処理が終了してない場合、ステップS31に戻って、新たなクラスチェーンを1つ取り出して、以上の処理を繰り返す。
【0098】次に、図14のステップS4に示すように、単語・トークンクラスタリング処理を行う。この単語・トークンクラスタリング処理では、テキストデータについての単語とトークンとの一次元列において、互いに異なる単語と互いに異なるトークンとを抽出し、単語とトークンとが混在する集合を単語・トークンクラス{T1 、T2 、T3 、T4 、・・・、TD }に分割する第2のクラスタリング処理を行う。
【0099】図21は、ステップS4の単語・トークンクラスタリング処理を示すフローチャートである。図21において、ステップS40に示すように、ステップS3で得られたテキストデータの単語・トークンの一次元列を入力データとして、ステップS1の第1の単語クラスタリング処理と同一の方法でクラスタリングを行うことより、単語・トークンクラス{T1 、T2 、T3 、T4 、・・・、TD }を生成する。この第2のクラスタリング処理では、単語とトークンは区別せず、トークンは1つの単語として扱われる。また、生成されたそれぞれの単語・トークンクラス{T1 、T2 、T3 、T4 、・・・、TD }は、その要素として単語とトークンを含んでいる。
【0100】次に、図14のステップS5に示すように、データ出力処理を行う。このデータ出力処理では、テキストデータの単語の一次元列に存在する単語列のうち、トークンに対応するものを連語として抽出し、単語・トークンクラス{T1 、T2、T3 、T4 、・・・、TD }の中のトークンを連語で置換することにより、単語と連語とが混在する集合を単語・連語クラス{R1 、R2 、R3 、R4 、・・・、RD }に分割する第3のクラスタリング処理を行う。
【0101】図22は、ステップS5のデータ出力処理を示すフローチャートである。図22において、まず、ステップS50に示すように、1つの単語・トークンクラスTi から1つのトークンtK を取り出す。
【0102】次に、ステップS51に示すように、テキストデータの単語の一次元列をスキャンし、ステップS52において、ステップS50で取り出したトークンtK に対応するクラスチェーンで結ばれた単語クラス列に属する単語列が存在するかどうかを判断する。そして、トークンtK に対応するクラスチェーンで結ばれた単語クラス列に属する単語列がテキストデータの単語の一次元列に存在する場合、ステップS53に進んで、この単語列を連語とみなす処理を繰り返し、テキストデータの単語の一次元列をスキャンすることにより得られたこれらの連語でトークンtK を置き換える。
【0103】一方、トークンtK に対応するクラスチェーンで結ばれた単語クラス列に属する単語列がテキストデータの単語の一次元列に存在しない場合、ステップS54に進んで、全てのトークンについて処理が終了したかどうかを判断し、全てのトークンについて処理が終了していない場合、ステップS50に進んで、以上の処理を繰り返す。
【0104】例えば、ステップS3のトークン置換処理において、テキストデータの単語の一次元列(w1 w2 w3 w4 ・・・ wT)のうち、単語列(w1 w2)、(w13w14)、・・・がトークンt1 で置換され、単語列(w4 w5 w6)、(w17w18)、・・・がトークンt2 で置換されたとすると、トークンt1 に対応する連語として、{w1 −w2 、w13−w14、・・・}がテキストデータから抽出され、トークンt2 に対応する連語として、{w4 −w5 −w6 、w17−w18、・・・}がテキストデータから抽出される。
【0105】1つの単語・トークンクラスTi が単語の集合Wi とトークンの集合Ji ={ti1、ti2、・・・tin}からなり、トークンクラスTi が{Wi UJi }により表され、、トークンの集合Ji の中の1つのトークンtim が、連語の集合Vim={vim(1) 、vim(2) 、・・・}に逆トークン置換されたとすると、1つの単語・連語クラスRi は、【0106】
【数2】

【0107】で与えられる。以上説明したように、本発明の一実施例による単語・連語分類処理装置によれば、単語と連語とを区別することなく分類することができる。
【0108】次に、本発明の一実施例による音声認識装置について説明する。図23は、図1の単語・連語分類処理装置により得られた単語・連語分類処理結果を利用して音声認識を行う音声認識装置の構成を示すブロック図である。
【0109】図23において、所定のテキストデータ40に含まれる単語と連語とが、単語・連語分類処理部41により単語と連語とが混在するクラスに分類され、この分類された単語と連語とが単語・連語辞書49に格納されている。
【0110】一方、複数の単語と連語とからなる発音音声は、マイクロフォン50によりアナログ音声信号に変換された後、A／D変換器51でデジタル音声信号に変換され、特徴抽出部52に入力される。特徴抽出部52は、デジタル音声信号に対して、例えば、LPC分析を行い、ケプストラム係数や対数パワーなどの特徴パラメータを抽出する。特徴抽出部52で抽出された特徴パラメータは、音声認識部54に出力され、音素隠れマルコフモデルなどの言語モデル55を参照するとともに、単語・連語辞書49に格納されている単語と連語との分類結果を参照しながら、単語及び連語ごとに音声認識を行う。
【0111】図24は、単語・連語分類処理結果を利用して音声認識を行う場合の例を示す図である。図24において、「本日は晴天なり」と発声された発音音声がマイクロフォン50に入力され、この発音音声に対して音声モデルを適用するとにより、例えば、「本日は晴天なり」という認識結果と「本日は静電なり」という認識結果とが得られる。これらの音声モデルによる認識結果に対し、言語モデルによる処理を行って単語・連語辞書49の参照を行い、「晴天なり」という連語が単語・連語辞書49に登録されている場合、「本日は晴天なり」という認識結果に対しては高い確率が与えられ、「本日は静電なり」という認識結果に対しては低い確率が与えられる。
【0112】以上説明したように、本発明の一実施例による音声認識装置によれば、単語・連語辞書49を参照して音声認識を行うことにより、より正確な認識処理が可能になる。
【0113】次に、本発明の一実施例による機械翻訳装置について説明する。図25は、図1の単語・連語分類処理装置により得られた単語・連語分類処理結果を利用して機械翻訳を行う機械翻訳装置の構成を示すブロック図である。
【0114】図25において、所定のテキストデータ40に含まれる単語と連語とが、単語・連語分類処理部41により単語と連語とが混在するクラスに分類され、この分類された単語と連語とが単語・連語辞書49に格納されている。また、用例原文とその用例原文に対する用例訳文とが、それぞれ対応させて用例文集60に格納されている。
【0115】用例検索部61に原文が入力されると、単語・連語辞書49を参照しながら入力された原文の単語が属するクラスを検索し、そのクラスと同一のクラスに属する単語又は連語により構成される用例原文を用例文集60から検索する。用例文集60から検索された用例原文及びその用例訳文は、用例適用部62に入力され、用例訳文の中の訳語を、入力された原文の単語に対する訳語に置換することにより、入力された原文に対する訳文を生成する。
【0116】図26は、単語・連語分類処理結果を利用して音声認識を行う場合の例を示す図である。図26において、"Toyota"と"Kohlberg Kravis Robert & Co. "とは同一のクラスに属し、"gained"と"lost"とは同一のクラスに属し、"2"と"1"とは同一のクラスに属し、"30 1／4"と"80 1／2"とは同一のクラスに属しているものとする。
【0117】原文として、"Toyota gained 2 to 30 1／4. "が入力されると、用例原文として、用例文集60から"Kohlberg Kravis Robert & Co. lost 1 to 80 1／2. "が検索されるとともに、その用例原文に対する用例訳文「Kohlberg Kravis Robert & Co. 社は、1ドル値を下げて終値80 1／2ドルだった。」も検索される。

【0118】次に、用例原文の原語"Kohlberg Kravis Robert &Co."と同一のクラスに属している入力原文の原語"Toyota"に対する訳語「トヨタ」で、用例訳文の訳語「Kohlberg Kravis Robert & Co.社」を置き換え、用例原文の原語"lost"と同一のクラスに属している入力原文の原語"gained"に対する訳語「上げて」で、用例訳文の訳語「下げて」を置き換え、用例訳文の数値"1"を"2"で置き換え、用例訳文の数値"80 1／2"を"30 1／4"で置き換えることにより、入力原文に対する訳文「トヨタは、2ドル値を上げて終値30 1／2ドルだった。」を出力する。

【0119】以上説明したように、本発明の一実施例による機械翻訳装置によれば、単語・連語辞書49を参照して機械翻訳を行うことにより、より正確な翻訳処理が可能になる。

【0120】以上、本発明の一実施例について説明したが、本発明は上述した実施例に限定されるものではなく、本発明の技術的思想の範囲内で他の様々な変更が可能である。例えば、上述した実施例では、単語・連語分類処理装置を音声認識装置及び機械翻訳装置に適用した場合について説明したが、単語・連語分類処理装置を文字認識装置に用いるようにしてもよい。また、上述した実施例では、単語と連語とを混在される分類する場合について説明したが、連語のみを抽出し、この抽出した連語を分類するようにしてもよい。

【0121】

【発明の効果】以上説明したように、本発明の単語・連語分類処理装置によれば、テキストデータに含まれる単語と連語とを一緒に分類して、単語と連語とが混在するクラスを生成することにより、単語と単語とをまとめて分類するだけでなく、単語と連語あるいは連語と連語とをまとめて分類することができ、単語と連語あるいは連語と連語との対応関係や類似度を容易に判別することができる。

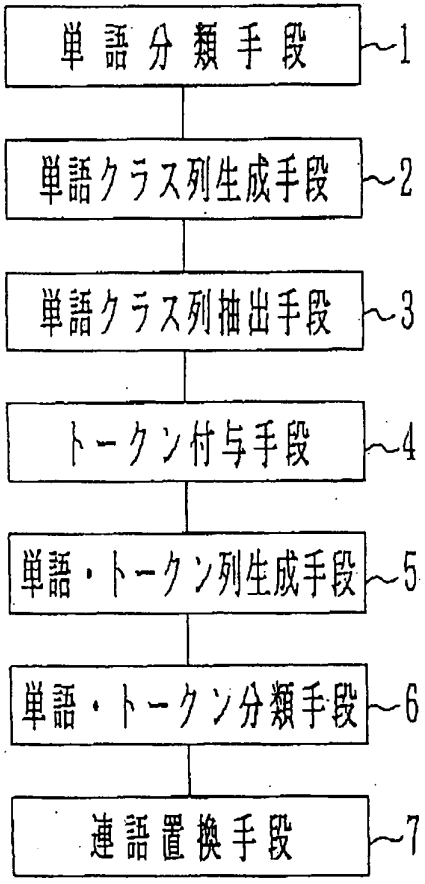
【0122】また、本発明の一態様によれば、テキストデータの単語クラス列にトークンを付与して単語クラス列を1つの単語とみなし、テキストデータに含まれる単語とトークンを付与された単語クラス列とを同等に取り扱ってこれらを分類してから、テキストデータに存在する単語列で対応する単語クラス列を置き換えるようにしたので、単語と連語との区別なく分類処理を行うことができるとともに、テキストデータからの連語の抽出を高速に行うことができる。

【0123】また、本発明の連語抽出装置によれば、テキストデータの単語列を構成する個々の単語を、その単語が属する単語クラスで置換し、テキストデータにおいて出現する確率が所定値以上の単語クラス列を抽出してから、テキストデータに存在する連語を抽出することにより、連語を高速に抽出することができる。

【0124】また、本発明の音声認識装置によれば、単語と連語あるいは連語と連語の対応関係や類似度を用いながら音声認識を行うことができ、正確な処理が可能になる。

【0125】また、本発明の機械翻訳装置によれば、用例文集に格納されている用例原文の単語が連語に置き換わった原文が入力された場合においても、入力された原文に用例原文を適用して機械翻訳を行うことができ、単語と連語あるいは連語と連語の対応関係や類似度を用いた正確な機械翻訳が可能になる。

本発明の一実施例に係わる単語・連語分類処理装置の機能的な構成を示すブロック図



本発明の一実施例に係わる単語・連語分類処理装置の単語クラスタリング処理を説明する図